Paper 125-28

The Use of Geographic Information Systems to Investigate Environmental Pollutants in Relationship to Medical Treatment

Patricia B. Cerrito, University of Louisville, Jewish Center for Advanced Medicine, Louisville, Kentucky
Robert Forbes, University of Louisville, Louisville, Kentucky
George R. Barnes, University of Louisville, Louisville, Kentucky

ABSTRACT

The purpose of this paper is to examine the use of geographic information systems (GIS) together with data mining techniques to examine the relationship between environmental exposure and the need to treat asthma in an emergency room setting. Patient data for those who were treated for asthma were collected in a database. This database contained approximately 1200 records including patient addresses. The addresses were mapped into GIS to compare to environmental hazards listed in the GIS database. In addition, address information from a lung cancer screening study was used to examine the relationship between geographic location and self-reporting of exposure to environmental hazards. Clustering, rule induction, and kernel density estimation were used to examine the data. It was found that African Americans represent 37% of the emergency room use while representing only 13% of the local population. In addition, African Americans have higher levels of exposure to environmental hazards while reporting lower levels, including exposure to cigarette smoke.

INTRODUCTION

This paper discusses the statistical methodology needed to combine environmental data with medical data by providing a specific example to investigate the relationship between environmental factors and the need for treatment of asthma and lung cancer. Two data sets will be used in this analysis. The first data set was collected over a two-year period and contains information about patients who were examined in the emergency room because they had shortness of breath. The second data set was also collected over a two-year period and includes information about patients requesting enrollment in a study of lung cancer. The second data set contains information concerning patient's perception of exposure to pollutants. The environmental information is stored in a Geographic Information System (GIS). The GIS can create maps linking medical data to environmental location to examine patterns and relationships in the data. To properly analyze the data, data mining tools must be used. Multiple levels of environmental factors combined with multiple patient factors add a level of complexity to the data that cannot be examined using standard statistical methods. Only data mining tools, designed to work with large, complex databases can examine the combined datasets.

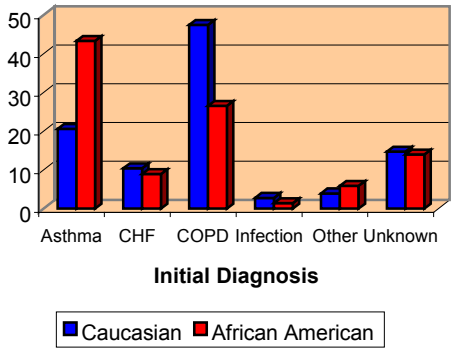
The GIS is a potent tool that has been under-utilized in medical studies. These studies indicate the potential of the GIS to investigate the relationship between environmental factors and the incidence of cancer. By using the GIS, it is possible to superimpose layers of environmental information. Each layer can be related to different environmental substances. Each value at a particular map point can be used as an explanatory variable in statistical models.

The primary finding in this study is that African Americans proportionally use the ER for treatment of shortness of air while routinely under-reporting exposure to cigarettes and other environmental factors. This under-reporting can contribute to less than optimal treatment in the African American population. In addition, patients at high risk in the ER for hospital admission because of shortness of air can be identified. This study also demonstrated the feasibility of using GIS in conjunction with patient data for determining risk factors,exposures, and demographic information in order to define populationsat risk for disease.

Asthma Database

The asthma data set has 1352 subjects and 126 variables. It contains the basic demographics for each patient, that is, age, race, height, weight, and the symptoms presented when the patient first came to the emergency room, as well as the emergency room diagnosis, and whether the patient was admitted to the hospital after ER treatment. A total of 460 out of 1206 patients reported are African Americans (38.14%, Figure 1), while 746 are Caucasians (61.86%). Note that the percentage of African Americans is substantially above the percentage residing in Louisville and Jefferson County (18%). The number of patients from other racial and ethnic groups was too small to be considered.

Figure 1. Comparison of ER Visits by Race



LUNG CANCER DATABASE

Data from a study involving lung cancer screening was used to examine COPD. Although 10.5% of the population recruited for the study were African American, only 3.5% were eligible for the study. There was a statistically significant difference in the two populations as noted in a second study of lung cancer that enrolled 1000 subjects after screening more than 3,545 for eligibility. It is interesting to note that African Americans reported smoking fewer cigarettes than did Caucasians (Table 1).

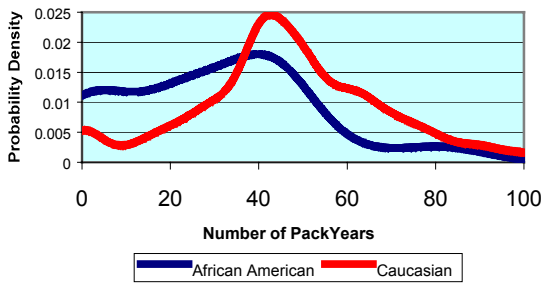
Table 1. Eligibility for Lung Cancer Screening

Race	Eligibility for Lung Cancer Study	Mean Number of Cigarettes Per Day	Mean Number of Pack Years	Mean FEV_1/FVC Ratio
Caucasian	No	18.02	35.51	68.37
African American	No	10.00	23.97	76.74
Caucasian	Yes	25.82	57.39	65.59
African American	Yes	21.00	49.39	56.98

In addition, only 8% of African Americans applying for enrollment into the lung cancer study reported emphysema compared to 15% of Caucasians who reported emphysema, although the rate of chronic bronchitis was 22.71% for both groups. The rate of asthma among African Americans applying for the screening

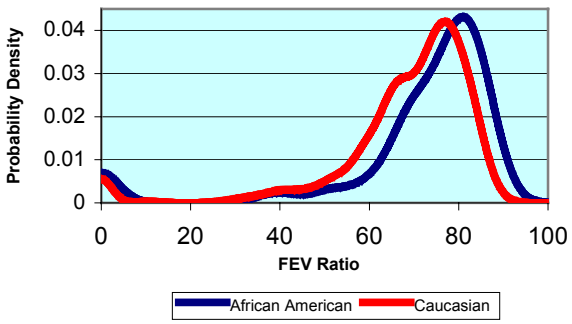
study was 18% compared to 14% for Caucasians. Figure 2 gives the distribution by race with for the reporting of the number of lifetime pack years.

Figure 2. Comparison of PackYears by Race



Overall, African Americans reported lower numbers of packyears than did Caucasians. This could be a problem in under-reporting or simply that African Americans smoke less. Similarly, shortness of air as detected by the forced expiratory volume (FEV) ratios differed between the populations (Figure 3).

Figure 3. Comparison of FEV Ratios by Race



FEV ratios are not subject to under-reporting. Therefore, it appears that there is less smoking in African Americans who have a higher rate of asthma. However, the rate of asthma reported (18%) still does not explain the high rate of emergency room visits (38%). Yet, as shown by the Cancer Atlas of the National Cancer Institute, African Americans have a much higher rate of mortality from lung cancer (Table 2).

Table 2. All cancers : White Males 1970 - 1994 [All Ages]

Location Name	Mortality Rate/100,000	Significantly Different from US *
Total US	209.47	---
Kentucky	225.14	Yes
Louisville, KY	253.16	Yes
Jefferson, KY	253.16	Yes

All cancers : Black Males 1970 - 1994 [All Ages]

Location Name	Mortality Rate/100,000	Significantly Different from US *
Total US	294.16	---
Kentucky	322.56	Yes
Louisville, KY	355.94	Yes
Jefferson, KY	355.94	Yes

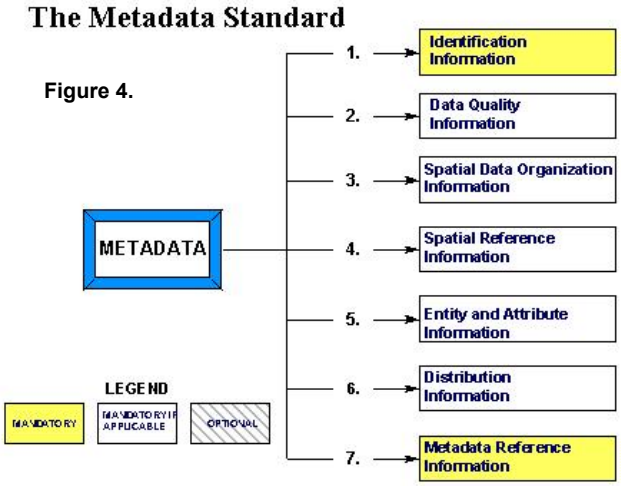
ANALYSIS AND GIS

Geographical Information System

At most, a handful of publications use the tool in epidemiological studies of cancer.[1-8] These few studies give an indication of the potential of GIS as a tool to investigate the relationship of environmental factors to the incidence of cancer.[9] With the GIS, it is possible to superimpose layers of information. Each value at a particular map point can be used as an explanatory variable in statistical models.[10]

In 1998, the Federal Geographic Data Committee approved the Content Standard for Digital Geospatial Metadata (FGDC-STD-001-1998). The standard consists of 7 sections of which 2 are mandatory:

Figure 4.



Spatial data contain coordinates and identifying information describing the map. Attribute data contain information that can be linked to the spatial data. It is necessary to have addresses that can be located on the map. According to the US Geological Survey:

> In the strictest sense, a GIS is a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information, i.e. data identified according to their locations. Practitioners also regard the total GIS as including operating personnel and the data that go into the system. With a GIS you can "point" at a location, object, or area on the screen and retrieve recorded information about it from off-screen files. A GIS can recognize and analyze the spatial relationships among mapped phenomena. Conditions of adjacency (what is next to what), containment (what is enclosed by what), and proximity (how close something is to something else) can be determined with a GIS. A GIS can simulate the route of materials along a linear network. It is possible to assign values such as direction and speed to the digital stream and "move" the contaminants through the stream system. A critical component of a GIS is its ability to produce graphics on the screen or on paper

that convey the results of analysis to the people who make decisions about resources.

GIS in SAS

The SAS software (SAS Institute, Inc.; Cary, NC) is the most sophisticated statistical tool generally available. It contains both data mining tools and GIS. There are a number of software packages that create GIS files. SAS can import many of them. In particular, it can import files in ARC/INFO format that is commonly used in many GIS centers. It can also import generic files provided that they identify X:east-west coordinate, Y-north-south coordinate, and an identifier. Additional attributes can be included. SAS uses a point, line, and area layer to construct a map. It can define a static map, or a thematic map with many layers where each layer represents themes by using different graphical characteristics. The advantage to using SAS for GIS construction is that the attribute data can be analyzed via statistical methods and data mining.

Statistical Methods

Little has been written concerning specific statistical methodologies to be used in conjunction with GIS.[3] Much of the analyses have focused on correlations.[1-8] However, there are statistical techniques developed with the express purpose of determining patterns in large datasets and examining causal flow from pollutant to asthma. Many of the tools used to examine spatial images in radiology can also be used to examine spatial images generated through GIS because they share many common characteristics.[11-14] Three data mining methodologies will be applied to the data: neural network analysis,[15] cluster analysis[16], and kernel density estimation.

Association rules are examined for X, a set of spatial variables and Y, set of health attributes such that X→Y with c% confidence. Since many such association rules will exist in the database, minimum support and minimum confidence are used. The support of a pattern A in a set of spatial objects S is the probability that a member of S satisfies pattern A and the confidence is the probability that pattern Y occurs if pattern X occurs where X→Y. The type of association rule to be examined in this project is

> Is_a(x,dry cleaning establishment) → close to (x, patient with asthma) with c% confidence?

The input consists of a spatial database, a mining query, and a set of thresholds as follows:

1. A database, that consists of three parts (1) a spatial database containing a set of spatial objects, (2) a relational database describing non-spatial properties of spatial objects, and a set of concept hierarchies;
2. A query that consists of (1) a reference set S of described objects, (2) task-relevant sets of spatial objects $C_1,\ldots,C_n$ that are used for description of the objects from S, and (3) a set of task-relevant spatial relations; and
3. Two thresholds for each level l of description: minimum support and minimum confidence interval.

The output is a strong multiple-level spatial association rule.

Definition 1. A spatial association rule is a rule in the form of $P_1 \wedge P_2 \wedge \ldots \wedge P_m \rightarrow Q_1 \wedge \ldots \wedge Q_n$ (c%) where at least one of the predicates is spatial and c% is the confidence of the rule which indicates that c% of objects satisfying the antecedent of the rule will also satisfy the consequent of the rule.

Definition 2. The support of a conjunction of predicates, $P=P_1 \wedge \ldots \wedge P_k$ in a set S, denoted as σ(P/S) is the number of objects in S which satisfy P versus the cardinality of S. The confidence of a rule P→Q in S, φ(P→Q/S), is the ratio of σ(P∧Q/S) versus σ(P/S). That is, the possibility that Q is satisfied by a member of S when P is satisfied by the same member of S. A single predicate is called 1-predicate. A conjunction of k single predicates is called a k-predicate.

Definition 3. A set of predicates P is large in set S at level k if the support of P is no less than its minimum support threshold σ_k' for level k, and all ancestors of P from the concept hierarchy are large at their corresponding levels. The confidence of a rule P→Q/S is high at level k if its confidence is no less than its corresponding minimum confidence threshold φ_k'.

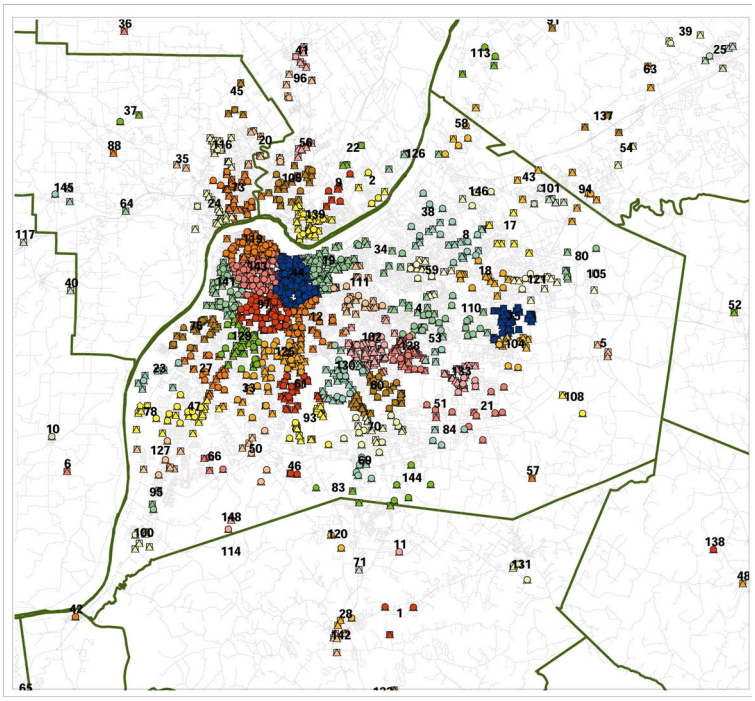
Definition 4. A rule P→Q/S is strong if predicate P∧Q is large in set S and the confidence of P→Q/S is high.

The X,Y coordinates identify the point location of objects. These coordinates can be used along with attribute variables to analyze data.

The initial cluster analysis done on the data failed because the addresses were too scattered. The database included patients visiting from out-of-town who had an asthma emergency. This caused the clusters to be too scattered to be of any real value.

A second analysis was done restricting the patient base to the local area. Different numbers of clusters were used; a maximum number of 150 was found to optimize outcomes (Figure 5).

Figure 5. GIS Map Indicating Locations of Environmentally Hazardous Sites



In addition to the cluster analysis, a neural network analysis was done to determine whether the patient addresses in the vicinity of waste sites could predict the level of patient risk asthma (while accounting for individual diagnosis). The result of the neural network was a 32% misclassification rate; 40% for a decision tree, and 50% for logistic regression.

Note that the cluster areas that have the largest proportion of hazardous waste sites are in the northwest area of Jefferson County, an area with a large population of African Americans. African Americans report a lower exposure to hazardous materials with a higher proportion not knowing their exposure (Table 3).

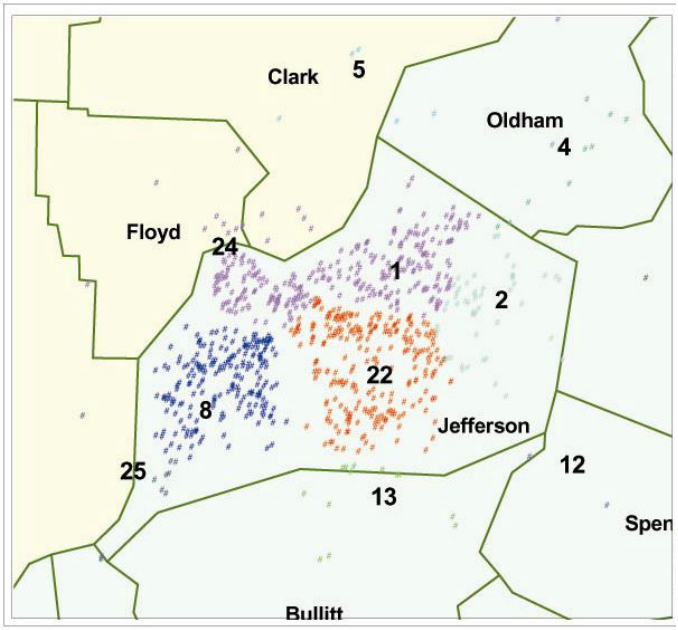
Table 3. Reporting of Exposure to Toxic Materials

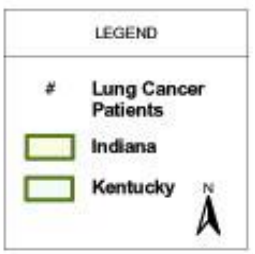
Exposure	African American	Caucasian	Not Known African American	Not Known Caucasian
Asbestos	11 (13.2%)	195 (16.8%)	26 (31.3%)	257 (22.1%)
Radon	4 (4.9%)	50 (4.3%)	19 (23.5%)	239 (20.8%)
Arsenic	2 (2.5%)	32 (2.8%)	20 (25.0%)	184 (16.1%)
Beryllium	0 (0%)	32 (2.8%)	21 (26.2%)	190 (16.6%)
Metals	3 (3.7%)	65 (5.7%)	23 (28.7%)	217 (18.9%)
Coal	1 (1.2%)	63 (5.5%)	21 (26.2%)	159 (13.8%)
Mustard Gas	1 (1.2%)	36 (3.1%)	19 (23.7%)	185 (16.2%)
Vinyl Chloride	1 (1.2%)	51 (4.5%)	22 (27.5%)	252 (22.1%)

Note that 31% of African Americans do not know whether they were exposed to asbestos compared to 22% of Caucasians. Similarly, Caucasians report almost five times as much exposure to coal compared to African Americans. Therefore, it appears that under-reporting in the African American Community is general in terms of exposure to environmental hazards, including cigarettes. Since most models that predict risk depend upon accurate patient information, African Americans will generally be perceived as having lower risk than they actually have.

As shown in the map, toxic sites are concentrated in the West End of Louisville. Therefore, residents in that area should report higher levels of exposure to the toxins in the end Table 2. The population involved in the lung cancer screening study was separated by clusters (Figure 6).

Figure 6. Patient Locations from Screening for Lung Cancer Study





Cluster 24, with the greatest proportion of African Americans in the Louisville area had a higher rate of smoking than any of the other clusters (Figure 7). However, the African Americans in cluster 24 reported a lower rate of current smoking (58.49%) than the didgeneral Caucasian population in the same cluster (Figure 8).

Figure 7. Proportion in Study Currently Smoking in Each Cluster

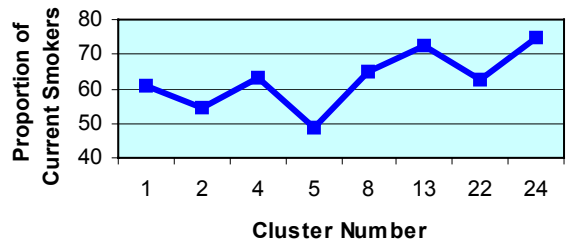
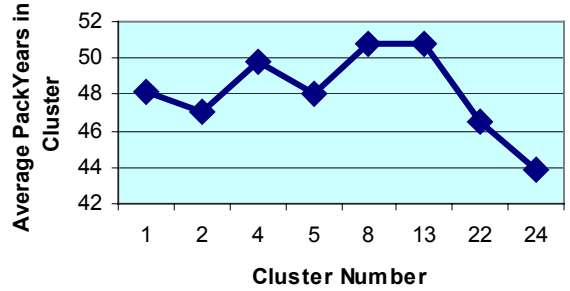


Figure 8. Average Lifetime Number of PackYears by Cluster



The discrepancy in reporting becomes clear when considering the National Cancer Institute's cancer mortality atlas. The rate of death from lung cancer is greater for African Americans than for Caucasians in Jefferson County (Table 4).

Table 4. Mortality from Lung Cancer in Jefferson County, 1990-1994

Population	Mortality from Cancer
Caucasian Male	101.84
African American Male	129.59
Caucasian Female	46.62
African American Female	53.7

Higher mortality rates from lung cancer seem to be related to higher rates of smoking.

Under-reporting cigarette use appears to be related to under-reporting exposure to environmental toxins. Map 1 indicates a higher concentration of toxins in cluster 24 compared to other areas in the local region. However, as Table 2 indicates, the reporting of exposure is similar or lower for African Americans. For cluster 24, the reporting is given in Table 5.

Table 5. Reporting of Exposure Restricted to Cluster 24

Exposure	Caucasian	African American	P-Value
Asbestos	23.13	11.32	0.1339
Radon	3.75	3.77	0.343
Arsenic	1.88	1.89	0.3858
Beryllium	3.13	0.00	0.1135
Metals	8.13	0.00	0.0228
Coal	5.00	0.00	0.019
Mustard Gas	2.50	0.00	0.1297
Vinyl Chloride	5.00	0.00	0.0494

Although asbestos is not statistically significant, due primarily to the small sample size in cluster 24, African Americans in the same geographic area reported half as much exposure to asbestos than did Caucasians.

CLASSIFICATION USING PATIENT DEMOGRAPHICS

Patient demographics were used to examine the likelihood that patients were admitted to the hospital with shortness of air after the ER visit.. The results are given in Table 6.

Table 6: Analysis of Maximum Likelihood (of being admitted to the hospital) Estimates of the Reduced Logistic Procedure

Parameter	Wald Chi-Square	Odds Ratio	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Pr > Chi-Square
SpO2	16.0565	3.65	0.994	1.937	6.875
Weight	3.7632	0.997	0.24	0.994	1
Cough Small	5.4221	0.461	0.159	0.24	0.885
Level (1 Versus 3)	29.9801	0.246	0.509	0.159	0.38
Race	4.0739	0.710	0.374	0.509	0.99
Modified diagnosis (Asthma Versus Other)	37.6615	0.579	2.295	0.374	0.896
Modified diagnosis (CHF Versus Other)	22.4971	4.565	2.295	2.295	9.081

The reduced logistic regression model was statistically significant in that the *p*-value for testing the null global hypothesis is less than 0.001. In addition, the c-statistic was 0.764, which indicates that the dependent variable is well explained by the model.

The odds ratio in the logistic regression gives the effect of one unit change in independent variable on the probability that the binary dependent variable equals 1. Therefore, an odds ratio greater than 1 indicates that the probability of being admitted to the hospital increases when the independent variable increases; an odds ratio less than 1 indicates that the probability of being admitted decreases when the independent variable increases. For example, SpO_2 has an odds ratio estimate of 3.65, which means that patients with unsaturated oxygen blood were 2.65 times more likely to be admitted to the hospital than those with saturated oxygen blood.

Those variables that have a 95% confidence interval of odds ratio estimated to be above 1 or below 1 were considered to contribute significantly to hospital admission. The variables SpO_2 and diagnosis of CHF have positive effects on the hospital admission in that the 95% confidence interval of odds ratio is above 1. So the patients who have unsaturated oxygen blood levels, and those diagnosed with CHF were more likely to be admitted to the hospital after ER treatment. The variables weight, presence of small cough, severity level, race (African American versus Caucasian), and asthma diagnosis versus other diagnoses have negative effects on whether the patient was admitted. That is, patients that had lowered acuity score, small cough and shortness of breath level 1 were less likely to be admitted, African American patients were less likely to be admitted, and patients diagnosed with asthma were less likely to be admitted. If the ER attending physician disregarded the standard protocol for treatment for shortness of breath also reduced the risk of hospital admission.

A second classification model was developed to determine whether patient demographics were related to patient location. A decision tree was used to examine classification (Figure 9). The variable that has the most discriminating power is race, followed by gender and smoking habits.

CONCLUSION

The highest level of exposure to toxins occurs in the West End of Louisville. However, this study suggests that there appears to be some consistent under-reporting of exposure by the population within that community.

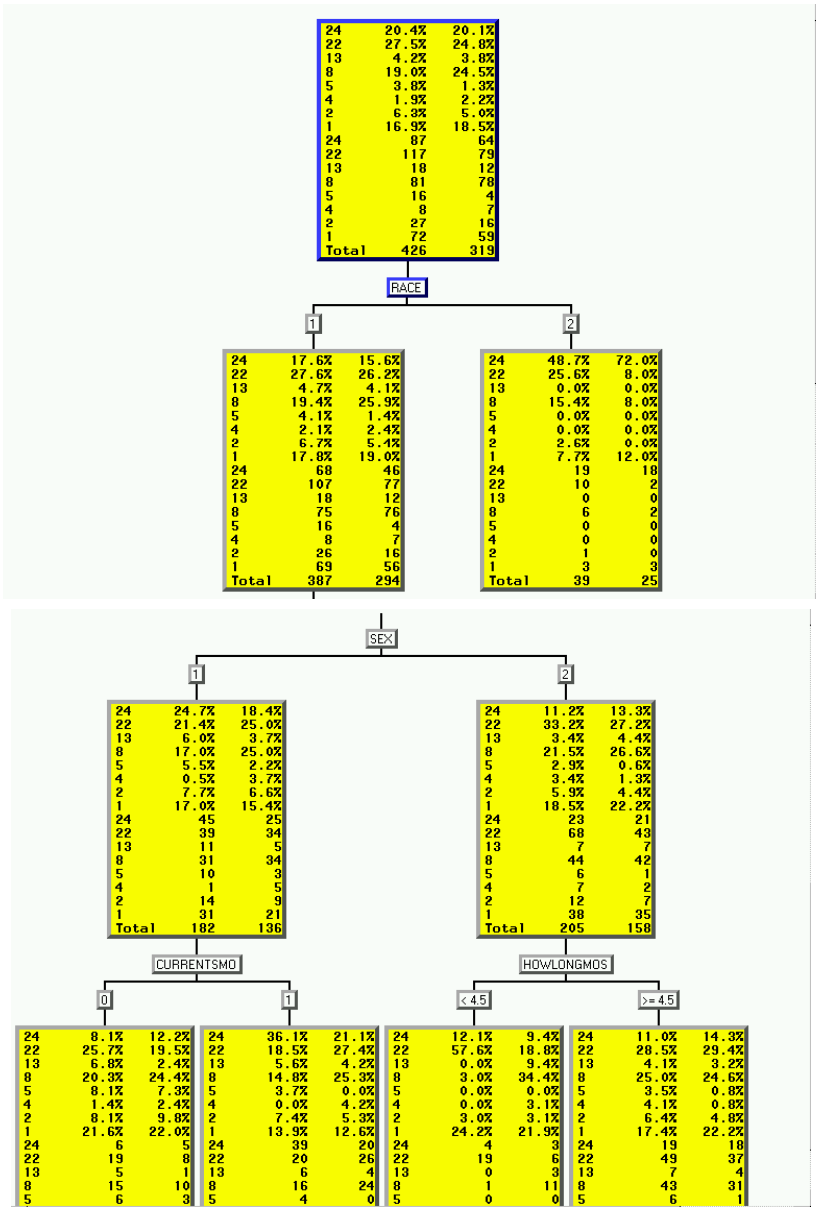
This study also demonstrated the feasibility of using GIS in conjunction with patient data for determining risk factors, exposures, and demographic information in order to define populations at risk for disease.

REFERENCES

1. White E, Aldrich TE. Geographic studies of pediatric cancer near hazardous waste sites. Arch Env Health. 54(6):390-397. 1999.
2. Harrison RM, Leung PL, Somervaille L, Smith R, Gilman E. Analysis of incidence of childhood cancer in the west midlands of the United Kingdom in relation to proximity to main roads and petrol stations. Occup & Env Health. 56(11):774-780. 1999.
3. Moskowitz, JM, Lin Z, Hudes ES. The impact of workplace smoking ordinances in California on smoking cessation. Am J Pub Health. 90(5):757-761. 2000.
4. Bickes JT. Community health assessment using computerized geographic mapping. Nurse Edu. 25(4):172,185. 2000.
5. Ramani de Silva S, Bundy ED, Smith PD, Gaydos .A Geographical Information System Technique for Record-Matching in a Study of Cancer Deaths in Welders. J Occupational and Environmental Med. 41(6)464-468. 1999.
6. Krautheim KR; Aldrich TE. Geographic information system (GIS) studies of cancer around NPL sites. Toxicology and Industrial Health.13(2-3): 357-62. 1997.
7. Hyndman JC; Holman CD. Differential effects on socioeconomic groups of modelling the location of mammography screening clinics using Geographic Information Systems. Australian and New Zealand Journal of Public Health. 24(3): 281-6. 2000.
8. Hjalmars U; Kulldorff M; Gustafsson G; Nagarwalla N. Childhood leukaemia in Sweden: using GIS and a spatial scan statistic for cluster detection. Statistics in Medicine. 15(7-9): 707-15. 1996.
9. Ward MH; Nuckols JR; Weigel SJ; Maxwell SK; Cantor KP; Miller RS. Identifying populations potentially exposed to agricultural pesticides using remote sensing and a Geographic Information System. Environmental Health Perspectives. 108(1): 5-12. 2000.
10. Spatial Analytical Methods and Geographic Information

Systems: Use in Health Research and Epidemiology. Epidemiologic Reviews. 21(2):143-161, 1999.
11. Dunn C, Kingham S. Establishing links between air quality and health: searching for the impossible? Soc Sci Med. 42(6):831-841. 1996.
12. Tourassi GD. Frederick ED. Vittitoe NF. Coleman RE. Fractal texture analysis of perfusion lung scans. Computers & Biomedical Research. 33(3):161-71, 2000.
13. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. Medical Decision Making. 17(2):186-92, 1997.
14. Tourassi GD. Floyd CE Jr. Artificial neural networks for single photon emission computed tomography. A study of cold lesion detection and localization. Investigative Radiology. 28(8):671-7, 1993.
15. Campbell NW. Thomas BT. Troscianko T. Automatic segmentation and classification of outdoor images using neural networks. International Journal of Neural Systems. 8(1):137-44, 1997.
16. Forgionne, Guisseppi A. Gangopadhyay, Aryya, Adya, Monica. Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems. Topics in Health Information Management. 21(1):21-34. 2000.

Figure 9. Decision Tree for Prediction of Patient Cluster Location Using Patient Demographics.



ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the Jewish Hospital Center for Advanced Medicine, 200 Abraham Flexner Way, Louisville, KY 40202

CONTACT INFORMATION

Patricia B. Cerrito[1,2], Robert Forbes[3], George Barnes[1]
[1] Department of Mathematics
[2] Jewish Hospital Center for Advanced Medicine
[3] Department of Geography and Geosciences
University of Louisville
Louisville, KY 40292

Work Phone: 502-560-8534
Fax: 502-852-7132
Email: pcerrito@louisville.edu